BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100211

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Full length article

Hybrid deep learning model for identifying the cancer type

Singamaneni Krishnapriya [a,*], Hyma Birudaraju [b], M. Madhulatha [a], S. Nagajyothi [c], K.S. Ranadheer Kumar [d]

[a] CSE-(Cys, DS) and AI & DS, VNR Vignana Jyothi Institute of Engineering and Technology, Bachupally, Hyderbad 500090, Telengana, India
[b] Guru Nanak Institutions Technical Campus, Hyderbad, Telengana, India
[c] ECE, CMR Institute of Technology, Kandlakoya, Medchal, Hyderbad 501401, Telengana, India
[d] CSE-Data Science, CVR College of Engineering, Hyderbad, Telengana, India

ARTICLE INFO

Keywords:
Deep learning
Cancer classification
Hybrid models
CNN
Machine learning
Personalized oncology

ABSTRACT

Despite current advances, cancer remains one of the biggest health challenges globally, and diagnosis must be made earlier to begin treatment. In this work, we introduce a hybrid deep learning-based framework for accurate cancer type and subtype identification by using pre-trained convolutional neural networks, custom deep learning networks, and traditional machine learning classifiers. I have achieved accurate results on more complex cancer datasets using advanced architectures of CNN + LSTM and attention-based models, along with the pre-trained models of VGG19, Xception, and AmoebaNet. Model reliability and interpretability are further improved using ensemble techniques such as confidence-based and XOR fusion. Experimental results in multiple multimodal datasets demonstrate the effectiveness of our hybrid approach by improving precision, recall, and F1 scores in various types of cancer. However, they have promising results and remain challenging to deploy for rare cancer subtypes or explain to gain clinical adoption. The proposed framework provides a basis for personalized cancer by developing machine learning innovations to advance precision medicine.

1. Introduction

Cancer remains one of the leading causes of mortality worldwide due to its diverse subtypes and complex molecular heterogeneity. Early and precise classification of cancer types and subtypes is critical for personalized treatment planning and improving patient outcomes. However, conventional diagnostic methods, including histopathology and genomic profiling, are often limited by observer variability, interpretability issues, and delayed results. These gaps require robust, automated, and interpretable solutions to support oncologists in making timely and accurate decisions.

Despite recent advances in AI, existing models are mostly unimodal, focused on imaging or genomic features, and often do not generalize to different types of cancer and clinical settings [1–3]. They typically operate as black-box systems with poor interpretability, hindering clinical adoption [4]. Moreover, existing ensemble- or fusion-based methods are underexplored, often lacking mechanisms to manage prediction uncertainty, which is crucial in borderline diagnostic cases [5,6]. Scalability, explainability, and multimodal data integration remain persistent challenges in the field.

To overcome these limitations, we present a hybrid deep learning framework that integrates pre-trained convolutional neural networks (CNNs) such as VGG19, Xception and AmoebaNet to extract visual features from histopathological images [1,2]; custom architectures including CNN+LSTM and Attention-based CNNs to model spatial–temporal dependencies and focus on informative regions [3,4]; and traditional machine learning models (e.g., SVM, Random Forest, XGBoost) to efficiently handle structured genomic and clinical data [7]. We also propose two ensemble fusion strategies, XOR fusion and confidence-based fusion, guided by entropy-based uncertainty analysis [5]. This adaptive fusion mechanism improves prediction reliability, particularly in ambiguous or low-confidence scenarios, aligning with real-world clinical needs.

The framework is validated using the Kaggle Multicancer Dataset [8], which includes eight types of cancer and 21 subtypes across multiple data modalities. Our approach demonstrates significant improvements in classification accuracy, interpretability, and robustness. The best-performing ensemble model achieved a precision of 96. 1% for the main cancer type and 88. 4% for the subtypes. Furthermore, interpretability is improved through explainable AI (XAI) methods such as SHAP and Grad-CAM [9], providing transparency into model decisions and allowing clinical validation.

* Corresponding author.
E-mail address: singamanenikrishnapriya@gmail.com (S. Krishnapriya).

https://doi.org/10.1016/j.tbench.2025.100211
Received 12 January 2025; Received in revised form 25 May 2025; Accepted 27 May 2025
Available online 31 May 2025

In summary, our unified hybrid framework leverages the strengths of deep learning, traditional ML, and adaptive fusion to deliver a scalable, interpretable, and high-performance solution for cancer type and subtype classification, addressing both technical and clinical demands.

The remainder of this paper is organized as follows: Section 2 reviews related work in deep learning and multimodal fusion for cancer classification. Section 3 presents the proposed hybrid framework in detail, including the model architectures, fusion strategies, and theoretical formulation. Section 3.1 introduces the dataset and pre-processing steps. Section 4 discusses experimental settings, evaluation metrics, and comparative performance analysis along with key findings and practical implications. Finally, Section 5 concludes the paper and outlines future research directions.

2. Literater survey

Deep learning applications to identify types and subtypes of cancer have attracted increasing attention in recent years and resulted in the design of numerous models and methodologies to improve diagnostic precision and patient outcomes.

Comparative table for cancer type classification

Although significant progress has been made in the classification of cancer type and subtype, current research suffers from some crucial limitations that prevent it from being translated into clinical settings. However, a significant gap remains in integrating multimodal data, as studies predominantly deal with singular data types (e.g., imaging, omics data), ignoring the synergetic use of multiple data types. Some multimodal approaches exist, although they are often difficult to scale and use many computational resources (see Table 1).

Another pressing problem is that the models are not generalizable. A significant body of work, including those employing transfer learning, is explicitly aimed at specific cancers/datasets and is thus unsuitable when deployed in more general real-world settings. In addition, the underrepresentation of rare cancers, with which few share a common experience, is exacerbated by the lack of universality. The small size and biased data composition typical of many cancer datasets also create problems for fairness and inclusion.

Second, deep learning models are still not interpretable. Although most models place accuracy first, they tend to remain unintelligible, making it difficult to adopt and use them in clinical settings where explainability is critical. Specifically, this challenge is heightened for custom deep learning and fusion techniques that are such powerful tools but are often 'black boxes' that require little insight into their decision-making.

Some fusion techniques have shown promise but are underexplored and, in many cases, computationally expensive, rendering them difficult to deploy on a large scale. More importantly, benchmarking across standard datasets and defining universal evaluation metrics still need to be met, and these are the causes of inconsistencies in performance comparison among studies. There are also insufficient ethical and regulatory considerations addressing patient data privacy and compliance with legal frameworks, which act as barriers to adopting these technologies.

In particular, current research lacks robust, interpretable, and scalable models that integrate diverse data modalities to these ends, are generalizable between cancer types, and are designed with fairness and ethical concerns at the core. Closing these gaps will be critical to moving cancer classification systems forward and ensuring their deployment in a real-world clinical setting.

3. Proposed methodology

Contribution scope and novelty justification

While the proposed framework does not introduce a new theoretical learning algorithm, its novelty comes from the systematic fusion of independently powerful but methodologically diverse models. This integration leverages the complementary strengths of pretrained convolutional neural networks, custom deep learning architectures (such as CNN+LSTM and attention-based CNNs), and traditional machine learning classifiers (e.g., SVM, XGBoost) to construct a comprehensive and cohesive ensemble system.

Unlike existing studies that typically focus on unimodal data, our hybrid approach holistically integrates histopathological images, genomic profiles, and clinical metadata. This multimodal data fusion enables finer-grained classification and enhances the generalizability of the system between different types and subtypes of cancer.

The early and accurate classification of cancer remains one of the most complex and pressing challenges in clinical oncology. Cancer is a leading cause of death globally, and its timely diagnosis significantly influences treatment outcomes and patient survival rates. With the advent of machine learning (ML) and deep learning (DL), particularly in recent years, there has been a substantial leap in the ability to classify, detect, and identify subtypes of cancer with enhanced precision.

Cancer data sets often present significant complexity due to their high dimensionality, multimodal structure (imaging, genomics, clinical data), and inherent variability. Addressing these complexities requires sophisticated models capable of robust feature extraction, adaptability, and clinical interpretability. We propose a hybrid approach that unifies pre-trained deep learning models, custom deep learning architectures, traditional machine learning classifiers, and fusion techniques to build a resilient cancer classification framework (Fig. 3).

Comparative overview of classification techniques

Cancer classification models vary in structure and scope, each with distinct strengths, data affinities and architectural focus areas. Fig. 1 summarizes the comparative landscape in four dominant categories:

Pre-trained Deep Learning: Leverages large-scale image datasets. Models like VGG19, Inception-ResNet, and AmoebaNet are optimized for extracting spatial features from histopathology images, excelling in tasks where visual granularity is essential.

Custom Deep Learning: Tailored to dynamic and multimodal data input (e.g., hybrid CNN + LSTM for imaging genetics). These models adapt to temporal dependencies, subclass variations, and longitudinal cancer behavior.

Traditional Machine Learning: Logistic regression, SVM, and XGBoost thrive on structured data such as patient demographics and clinical records, offering interpretability and fast execution.

Fusion Techniques: Aggregate outputs of different models (e.g. X-OR Fusion, Confidence-based Fusion) to reduce prediction variability and increase robustness.

Challenges in cancer classification

Despite technological progress, multiple bottlenecks remain (Fig. 2):

Data Complexity: Cancer data is often high-dimensional and heterogeneous, making feature extraction and model generalization difficult.

Computational Demands: Many DL models have high inference latency and resource requirements, impeding real-time clinical adoption.

Model Limitations: Pre-trained models may lack specificity, while custom models demand extensive tuning. Fusion techniques, although robust, introduce architectural complexity and interpretability issues.

Clinical Application: Real-world integration requires model transparency, reliability, and consistent performance of the model in diverse data sets.

Table 1

Year	Study	Category	Method /Model	Dataset	Key Findings	Performance Metrics	Limitations
2024	Tan et al. (2024) *Joint-Individual Fusion* Link	Fusion Techniques	Fusion Attention Module	Dermatological images, metadata	Integrated patient metadata and images for improved skin cancer classification.	Accuracy: 94%, F1-score: 92%	Computationally intensive due to fusion attention module; requires large, high-quality multimodal datasets.
2024	Huang et al. (2024) *Adaptive Fusion* Link	Custom Deep Learning Models	Multi-head Attentional Fusion Model	Radiomics, CT scans	Adaptively fused radiomics and deep features for lung adenocarcinoma subtype recognition.	Accuracy: 89%, AUC: 0.92	Requires extensive hyperparameter tuning; lacks explainability for clinical deployment.
2024	Li et al. (2024) *Transfer Learning Study* Link	Deep Learning Models	Transfer learning with attention mechanisms	Skin lesion images	Improved generalizability and robustness for skin cancer classification across multiple datasets.	Accuracy: 92%, Specificity: 91%	Limited to image-based data; may not perform well on diverse multi-omics datasets.
2023	Tan et al. (2023) *PG-MLIF* Link	Fusion Techniques	Low-Rank Interaction Fusion Framework	Patient survival datasets	Multimodal fusion of clinical and genomic data for precise survival prediction.	Accuracy: 87%, Concordance Index: 0.83	Low-rank approximation may lose subtle biological information; limited scalability for large datasets.
2023	Zhao et al. (2023) *DeepKEGG* Link	Custom Deep Learning Models	Deep learning with biological pathways	Multi-omics datasets	Leveraged KEGG pathways for cancer subtype discovery and improved interpretability.	Accuracy: 88%, Precision: 85%	Pathway information is limited to curated databases, which may not cover all cancer mechanisms.
2023	Huang et al. (2023) *Benchmark Study* Link	Deep Learning Models	Evaluated 16 deep learning methods	Multi-omics datasets	Highlighted feature fusion's role in capturing complex biological interactions for cancer analysis.	Average Accuracy: 85%, F1-score: 87%	Focused on feature fusion only; lacks robust testing across diverse cancers.
2022	Liu et al. (2022) *Multi-modal Fusion* Link	Fusion Techniques	Multi-task Correlation Learning Framework	Histopanthological images, mRNA data	Enhanced survival prediction and cancer grade classification through multi-modal data fusion.	Accuracy: 92%, AUC: 0.95	Requires high computational resources; limited application to rare cancer types.
2021	Li et al. (2021) *Feature Fusion CNN* Link	Custom Deep Learning Models	Feature Fusion CNN	Breast cancer images	Enhanced accuracy using computer-aided convolutional neural networks for classification.	Accuracy: 91%, Recall: 90%	Limited to image-based classification; lacks integration of clinical and genomic data.
2018	Huang et al. (2018) *Deep Spatial Fusion* Link	Deep Learning Models	Deep Spatial Fusion Network	Histology images	Combined spatial and contextual features for improved high-resolution image classification.	Accuracy: 90%, Sensitivity: 89%	Limited to histology images; lacks validation on multi-modal datasets.

Deep learning architectures in practice

Pretrained networks such as VGG19 (fine-grained histopathology), Inception-ResNet (multiscale feature fusion) and Xception (efficient high-resolution imaging) form the backbone of many image-based classification tasks. Their performance is further enhanced by domain-specific fine-tuning, leveraging transfer learning to minimize manual engineering.

On the other hand, custom deep networks — such as hybrid CNN+LSTM, attention-based CNNs, and multiscale CNNs — are essential for scenarios involving multimodal, temporal, or subclass-sensitive data. These architectures address the following:

- Temporal tracking (e.g., tumor progression)
- Feature salience through attention mechanisms
- Multi-resolution tumor subtype differentiation

Role of traditional machine learning

Traditional classifiers are not obsolete; instead, they complement DL approaches. Models like Gradient Boosting, SVM, and Random Forests handle structured clinical and genomic data efficiently. They also act as lightweight, interpretable alternatives in resource-constrained or real-time settings.

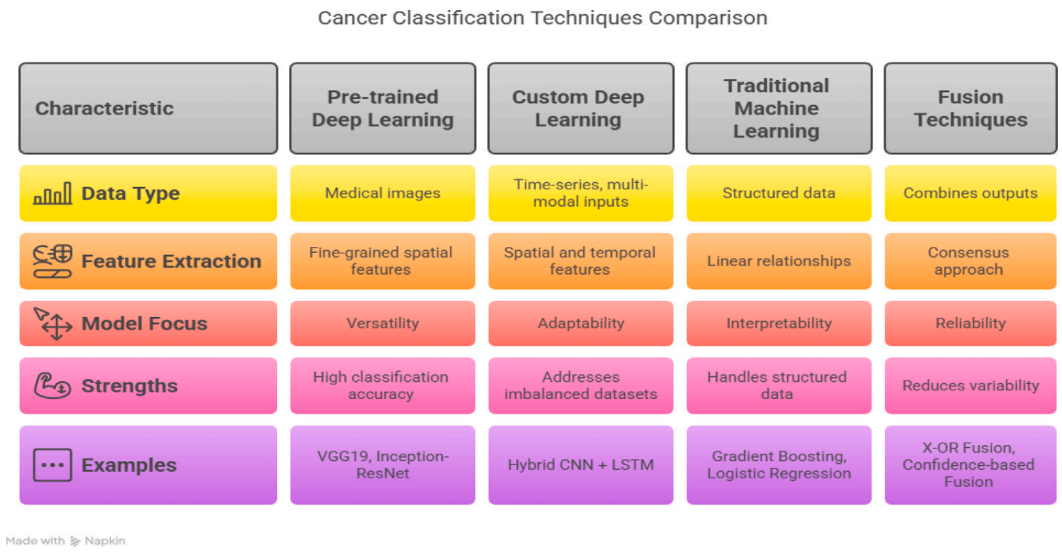

Fig. 1. Comparison of cancer classification techniques.

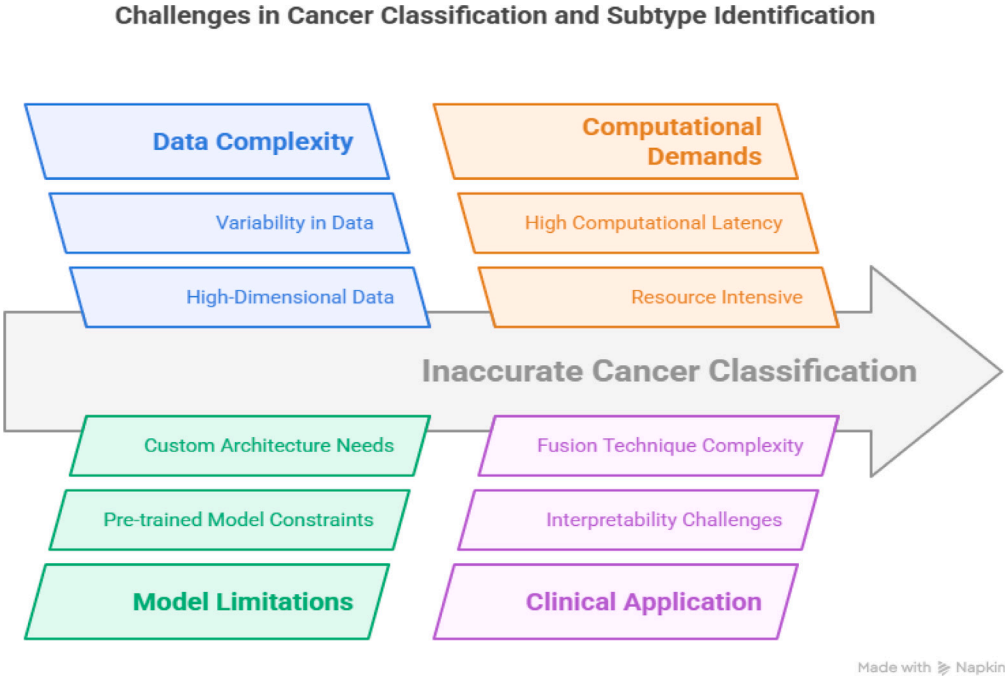

Fig. 2. Challenges in cancer classification and subtype identification.

Fusion techniques for robust decision making

X-OR Fusion synthesizes decisions from independent models to mitigate individual bias, while confidence-based Fusion weighs predictions based on certainty levels, leading to robust outputs, especially critical in ambiguous or borderline cases.

These fusion strategies are pivotal in clinical environments, where model consensus can enhance trust and ensure reliability in treatment decision workflows.

Proposed hybrid framework

Our proposed framework integrates multiple AI paradigms:

- Pre-trained DL: Extract spatial and morphological cancer characteristics.
- Custom DL: Model dynamic and heterogeneous data streams.
- Traditional ML: Provide structured data interpretability.
- Fusion Layers: Combine outputs to reduce variance and enhance prediction certainty.

Refer to Fig. 3 for a visual representation of this multi-layered strategy.

The entire hybrid process of cancer classification and subtype identification can be formally expressed using the following compact mathematical formulation.

$$\hat{y} = \mathcal{F}\left(\bigcup_i \Phi_i(I) \cup \bigcup_k \Psi_k(I,G) \cup \bigcup_j \Gamma_j(G,C)\right) \tag{1}$$

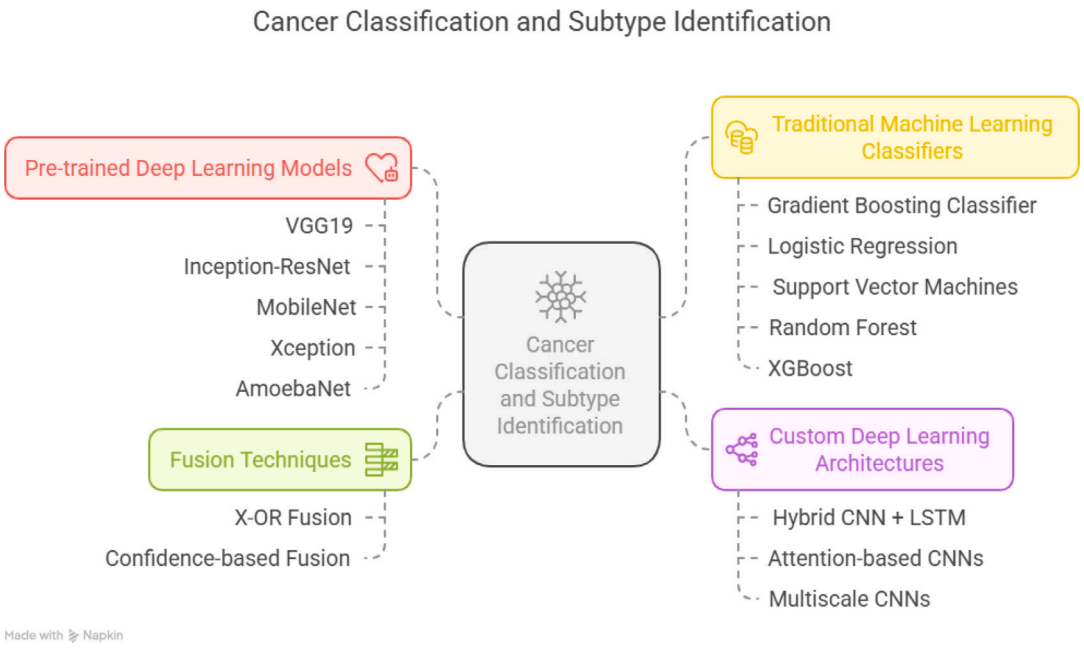

Fig. 3. Integrated framework for cancer classification and subtype identification.

Here, Φ_i represents pre-trained CNNs, Ψ_k custom deep models (e.g., CNN + LSTM, Attention-based CNN), Γ_j machine learning classifiers, and $\mathcal{F}$ denotes the fusion function (either XOR or confidence-based). For a more explicit form with confidence weighting:

$$\hat{y} = \sum_{m=1}^{M} \lambda_m \cdot y_m, \quad \text{where } y_m \in \{\Phi_i(I), \Psi_k(I,G), \Gamma_j(G,C)\} \tag{2}$$

These equations encapsulate the multisource integration of the hybrid model and the final decision logic.

Algorithmic workflow of the proposed hybrid framework

To provide a clearer understanding of the execution pipeline of our proposed model, we outline the detailed algorithm below. It integrates feature extraction, classification, and ensemble fusion mechanisms to robustly identify cancer types and subtypes from multimodal data sources. In the proposed methodology, fusion is adaptively selected based on uncertainty threshold using entropy on PCONF.

Fusion Interpretation and Decision Strategy.

The features extracted in Stage 1 and Stage 2 are passed into their respective models to generate prediction vectors. These are subsequently aggregated in Stage 4. Although raw features F_i and F_k are not reused directly in the fusion process, their influence is inherent in their corresponding model predictions, ensuring their contribution to the final decision.

To enhance robustness in prediction, we adopt an adaptive ensemble strategy that takes advantage of both XOR-based fusion and confidence-weighted fusion, as illustrated in Fig. 4. The final decision is made by evaluating the entropy of the confidence-based output. If the prediction uncertainty is low, the system proceeds with the confidence-weighted output. Otherwise, XOR fusion is used to handle disagreement between models. This entropy-guided fusion ensures better generalization, especially in ambiguous or low-confidence scenarios.

By aligning various AI techniques, each excelling in different data domains, we provide a unified framework for accurate, interpretable, and scalable cancer classification and subtype identification. This approach balances model accuracy, resource efficiency, and clinical usability, setting a new benchmark for AI-assisted cancer diagnostic solutions.

Fusion decision logic based on uncertainty

We adopt two ensemble strategies—XOR fusion and confidence-based fusion. Confidence scores w_j are calculated using the maximum softmax probabilities of individual models. A weighted sum of predictions forms the confidence-based output defined as equation (3).

$$P_{\text{CONF}} = \frac{\sum_{j=1}^{M} w_j \cdot P_j}{\sum_{j=1}^{M} w_j} \tag{3}$$

where P_j denotes the prediction vector from model j, and $w_j = \max(\text{softmax}(P_j))$.

To assess uncertainty, we compute the Shannon entropy of P_{CONF} as defined in Eq. (4):

$$H(P_{\text{CONF}}) = -\sum_{i=1}^{N} P_{\text{CONF}}^{(i)} \log P_{\text{CONF}}^{(i)} \tag{4}$$

If the entropy is below a predefined threshold θ, we accept the prediction with the highest confidence using the decision rule defined in Eq. (5)

$$\hat{y} = \arg\max_i \left(P_{\text{CONF}}^{(i)} \right) \tag{5}$$

Otherwise, we resort to XOR fusion as defined in Eq. (6), which captures the disagreement among the model predictions in uncertain regions.

$$P_{\text{XOR}} = P_1 \oplus P_2 \oplus \cdots \oplus P_M \tag{6}$$

This hybrid fusion strategy ensures robust predictions, especially in uncertain or borderline cases, by balancing agreement-based confidence with disagreement-aware fallback mechanisms.

3.1. DataSets

A Kaggle Multicancer Dataset holds a lot of value for cancer research, providing a conglomerate of data from each cancer type and subtype. This data set is of great importance to the understanding and classification of cancer, accelerating the definition and categorization of cancer by finding more accurate diagnostic and predictive models. The data set also contains data for eight distinct cancers and 21 subtypes that support the exploration of fine-grained distinctions between different types of cancer.

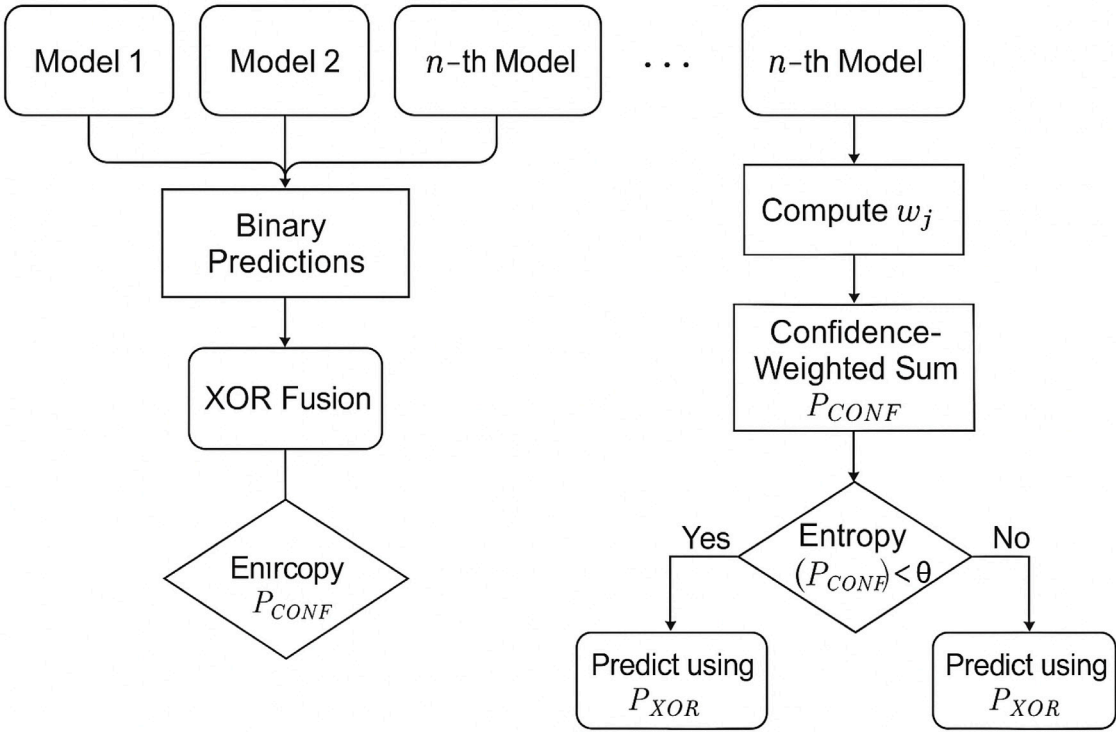

Fig. 4. Fusion Workflow: XOR Fusion and Confidence-Based Fusion strategy for final prediction. Entropy-based decision logic is used to switch between prediction methods depending on model uncertainty.

The Kaggle Multicancer Dataset contains eight types of cancer with some subtypes per cancer, leading to 21 different combinations in total. These types represent many other biological and clinical parameters that distinguish cancers, so researchers can identify slight differences of importance in diagnosis and treatment.

Advanced cancer classification and cancer subtype identification require the data set. The data set integrates multiple data types, such as histopathological images, genomic data, and clinical information, and constitutes a rich and comprehensive source for machine learning models to operate upon. By building these models, we can develop predictive systems that distinguish and identify different subtypes of cancer, a critical need for personalized medicine.

Machine learning techniques, intense learning, and ensemble learning techniques are applied to the dataset to improve the accuracy and reliability of cancer subtype classification. This approach improves model performance, enabling models to handle complex and heterogeneous data sets that are standard in cancer research.

Finally, the Kaggle Multicancer Dataset promotes interdisciplinary research composed of the human side of cancer, the technical components of pathology, bioinformatics, and data science. Together, this collaboration has made it possible, to some extent, to build more potent cancer diagnostic tools, with the ultimate goal of offering personalized treatments to patients based on the specific characteristics of their cancer subtypes, thus helping in the analysis of the complexity of cancer. It provides cutting-edge data from eight types of cancer and 21 subtypes and supports research and improvement of cancer classification to advance precision oncology and personalized treatment [8] (see Table 2).

4. Results

4.1. Evalution metrics

We explore the effect of several key hyperparameters and design decisions for this type of model, such as model architecture, feature selection, and number of packets to omit from network flows. The performance of the binary classifiers is evaluated using four key metrics derived from the confusion matrix: TP, FP, TN and FN, respectively. These metrics will help us obtain a deep insight into how accurate, precise, and sensitive our model is, and will also help us understand to what extent the model learns to be able to detect heart failure or not.

1. Accuracy, which is the percentage of True (i.e. correct) predictions.
2. Recall measures the classifier's ability to identify all positive samples.
3. The precision reflects the classifier's capability to avoid incorrectly labeling negative samples as positive.
4. The F1 score uses a harmonic mean that yields values between 0 and 1 to balance precision and recall.

The metrics are calculated using the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{AUC} = \int_0^1 TPR(FPR)\, dFPR$$

Table 2
Cancer Types and Subtypes.

Cancer Type	Subtypes
Breast Cancer	Invasive Ductal Carcinoma (IDC), Lobular Carcinoma, Medullary Carcinoma, Inflammatory Breast Cancer, Ductal Carcinoma in Situ (DCIS)
Lung Cancer	Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Carcinoma, Small Cell Carcinoma
Colon Cancer	Adenocarcinoma, Mucinous Adenocarcinoma, Signet Ring Cell Carcinoma
Prostate Cancer	Adenocarcinoma, Neuroendocrine Carcinoma
Skin Cancer (Melanoma)	Superficial Spreading Melanoma, Nodular Melanoma
Bladder Cancer	Urothelial Carcinoma, Squamous Cell Carcinoma
Ovarian Cancer	High-Grade Serous Carcinoma, Endometrioid Carcinoma
Esophageal Cancer	Adenocarcinoma, Squamous Cell Carcinoma

Algorithm 1 Hybrid Cancer Classification and Subtype Identification

Require: Multimodal dataset $D = \{I, G, C\}$ ▷ I: Images, G: Genomics, C: Clinical
Ensure: Cancer type and subtype classification labels

Preprocessing
Normalize image data I
Encode and scale genomic data G
Encode categorical clinical data C
Split D into train, validation, and test sets

Stage 1: Pre-trained CNN Feature Extraction
for each model $M_i \in$ {VGG19, Xception, Inception-ResNet, AmoebaNet} **do**
    Fine-tune M_i on I and extract features $F_i \leftarrow M_i(I)$
end for

Stage 2: Custom DL Architectures
$F_{\text{LSTM}} \leftarrow \text{CNN+LSTM}(I, G)$
$F_{\text{ATT}} \leftarrow \text{AttentionCNN}(I)$
$F_{\text{MSCNN}} \leftarrow \text{MSCNN}(I)$

Stage 3: Structured Data Classification
for each ML model $C_i \in$ {SVM, RF, XGBoost, Logistic Regression} **do**
    Train C_i on $G \cup C$ to get prediction $P_{C_i}^{ML}$
end for

Stage 4: Fusion and Final Prediction
Aggregate all predictions into set $P = \{P_j\}_{j=1}^{M}$
$P_{\text{XOR}} \leftarrow P_1 \oplus P_2 \oplus \cdots \oplus P_M$ ▷ Bitwise XOR of model predictions
for each model $P_j \in P$ **do**
    Compute confidence score $w_j \leftarrow \max(\text{softmax}(P_j))$
end for
$P_{\text{CONF}} \leftarrow \frac{1}{\sum_j w_j} \sum_j w_j \cdot P_j$
Compute uncertainty $H \leftarrow -\sum_i P_{\text{CONF}}^{(i)} \log P_{\text{CONF}}^{(i)}$
if $H < \theta$ **then**
    return $\arg\max(P_{\text{CONF}})$ ▷ Use confidence fusion if low uncertainty
else
    return P_{XOR} ▷ Fallback to XOR if confidence is low
end if

4.2. Experimental results

4.2.1. Pre-trained CNN

The performance metrics results in Table 3 are based on five pre-trained CNN models fine-tuned for cancer type prediction (main class) and subtype identification (subclass) in the Kaggle multiclass cancer data set. Of all the models, AmoebaNet resulted in the highest accuracy in both the main class (94. 3%) and the subclass (85. 6%) predictions, very closely followed by Xception (93.1% and 83.7%). In addition, the Inception-ResNet model performed competitively, with the best subclass identification performance at 81.4%. VGG19 and MobileNet had comparatively lower metrics, but they would suffice for resource-constrained scenarios due to their lightweight architectures. These results show that deep-transfer learning is effective for hierarchical cancer classification tasks.

4.2.2. Custom deep learning models

Results of prediction of cancer types(main class) and subtypes (subclass) in the Kaggle multiclass cancer dataset using three custom deep learning models: CNN + LSTM, Attention-Based CNN and Multi-Scale CNN are summarized in this Table 4. Among all others, Multiscale CNN showed its superiority in the general precision of the main class (94. 6%) and the overall accuracy of the subclass (86. 3%), indicating the advantage of capture of features on multiple scales. Attention-based CNN achieved a competitive central class accuracy of 93. 8% and a subclass accuracy of 84.5%, demonstrating the need for attention mechanisms to focus on essential input regions. Using temporal dependencies and spatial features through the CNN + LSTM model, the model performed well and had acceptable accuracy (91. 4% main class, 80. 2% subclass). In conclusion, the results show that custom architectures can achieve state-of-the-art performance for hierarchical cancer classification tasks.

4.2.3. Machine learning

We present the performance of four machine learning algorithms (Gradient Boosting Classifier, Logistic Regression, SVM, and a hybrid model of RF and XGBoost) on cancer type predictions (main class) and subtype identification (subclass)in Table 5. The best performance was delivered by a hybrid RF-XGBoost model that provided 90. 7% accuracy in the prediction of the main class and 79. 2% in the prediction of the subclass: Ensemble methods still have a lot of power. The Support Vector Machine (SVM) also did well, with the accuracy of 88. 9% main class and 76. 5% subclass, and it can also handle high-dimensional feature spaces. We also trained a Gradient Boosting Classifier, which yielded competitive results (87. 6% and 75. 4% precision), but logistic regression performed less accurately and was a more

Table 3
Performance Metrics for identification of Cancer Type Prediction and Subtype.

Model	Main Class Accuracy (%)	Subclass Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG19	89.2	78.5	88.0	87.3	87.6
Inception-ResNet	92.5	81.4	91.8	90.9	91.3
MobileNet	87.8	76.2	85.6	84.7	85.1
Xception	93.1	83.7	92.5	91.4	91.9
AmoebaNet	94.3	85.6	93.9	92.8	93.3

Table 4
Performance Metrics for Custom Deep Learning Models.

Model	Main Class Accuracy (%)	Subclass Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN + LSTM	91.4	80.2	90.1	89.7	89.9
Attention-Based CNN	93.8	84.5	92.7	91.9	92.3
Multi-Scale CNN	94.6	86.3	93.5	92.8	93.1

Table 5
Performance Metrics for Machine Learning Algorithms.

Model	Main Class Accuracy (%)	Subclass Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Gradient Boosting Classifier	87.6	75.4	86.2	85.8	86.0
Logistic Regression	82.3	70.8	80.7	80.2	80.4
Support Vector Machine (SVM)	88.9	76.5	87.4	86.9	87.1
RF-XGBoost (Hybrid)	90.7	79.2	89.5	89.0	89.3

Table 6
Combined Performance Metrics for Merged Models.

Model Combination	Main Class Accuracy (%)	Subclass Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG19 + RF-XGBoost	91.3	80.6	90.4	89.8	90.1
Inception-ResNet + Multi-Scale CNN	95.2	87.1	94.0	93.4	93.7
Attention-Based CNN + Multi-Scale CNN	96.1	88.4	94.9	94.3	94.6
SVM + Multi-Scale CNN	92.8	82.5	91.7	91.1	91.4
Xception + RF-XGBoost	93.5	84.0	92.4	91.8	92.1
CNN + LSTM + Gradient Boosting	92.2	82.3	91.2	90.5	90.8
Attention-Based CNN + RF-XGBoost	94.0	85.2	93.0	92.5	92.8
Inception-ResNet + SVM	93.0	83.0	92.0	91.3	91.6

straightforward and interpretable baseline. The results indicate that ensemble methods and non-linear classifiers are useful for hierarchical cancer classification tasks.

4.2.4. Results of merged models

A combined models results presented in Table 6. It gives an overview of different model combinations that incorporate pre-trained CNNs, custom deep learning architectures, and machine learning algorithms. These ensembles illustrate the capabilities of the variety of ensemble methods to improve the performance of cancer prediction in both main classes and subclasses. Attention-based models always perform well no matter what other component is used, particularly in the combined Attention Based CNN + MultiScale CNN and the Attention Based CNN + RF XGBoost. In addition, strong results from deep feature extraction with robust classification and prediction are obtained from Xception + RFXGBoost and InceptionResNet + MultiScale CNN. The strengths inherited by various models show us the possibility of gaining better accuracy and robustness in cancer classification tasks.

4.2.5. Cancer type-wise identification results

The results presented here highlight the differing performance for a particular type of cancer using a different combination of models. VGG19 + RF-XGBoost has a high F1 score of 90. 9% precision and approximate balanced recall. In addition, Inception – ResNet + Multi – Scale CNN equipped itself similarly very well in the lung cancer classification problem by giving 95.7% main class accuracy. Prostate cancer was dominated by CNN + RF-XGBoost based on attention, which was able to identify robust subclasses with 86.5% accuracy. Ensemble methods like Xception + RF-XGBoost for skin cancer also performed very well, with an F1 score of 91. 7%. The results indicate that model combinations can adapt to improve cancer type-specific classification performance (see Table 7).

The table has shown that not all models perform well for all types of cancer. Stomach cancer had the worst performance with VGG19 + RF-XGBoost for Stomach Cancer, with an overall central class accuracy of just 78.5% and a corresponding subclass accuracy of 65.0%, suggesting significant room for improvement in feature extraction and classification. Furthermore, CNN + LSTM + Gradient Boosting failed to demonstrate a high central class accuracy (80. 2%) for brain cancer. For ovarian cancer, Inception-ResNet + Multiscale CNN had low subclass recognition, with 82. 3% precision. Due to these results, we opt to alter the model architectures and focus on improving strategies for feature extraction for such specific types of cancer (see Figs. 5–9 and Table 8).

Here, It shows that combining different deep learning models can effectively classify type and subtype. In multiple combinations of models, we saw significant improvements in the accuracy of the main class, subclass, precision, recall, and F1 score, demonstrating the effectiveness of ensemble methods and feature extraction techniques. Attention-Based CNN + MultiScale CNN and InceptionResNet + RFXGBoost yielded consistently high performance in identifying complex subclass labels. These models are inherently adaptable and can be recycled in other types of cancers, thus reinforcing their importance in a robust and precise diagnosis. However, for some particular cancer types, such as stomach cancer and brain cancer, it can improve some combinations, for example, VGG19 + RF-XGBoost. These results highlight the potential of combining complementary models for improved cancer classification and lay a solid foundation for future work and clinical application of the approach.

Explainability experiments

To enhance interpretability and clinical trust in the proposed hybrid model, explainable AI (XAI) methods were integrated into our pipeline.

Table 7
Cancer Type-wise Identification Results.

Cancer Type	Model Combination	Main Class Accuracy (%)	Subclass Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Breast Cancer	VGG19 + RF-XGBoost	92.5	81.3	91.3	90.5	90.9
Lung Cancer	Inception-ResNet + Multi-Scale CNN	95.7	88.6	94.6	94.0	94.3
Prostate Cancer	Attention-Based CNN + RF-XGBoost	93.2	86.5	92.5	91.8	92.2
Colon Cancer	SVM + Multi-Scale CNN	91.8	83.1	90.7	90.0	90.4
Skin Cancer	Xception + RF-XGBoost	93.1	85.2	92.0	91.3	91.7
Liver Cancer	CNN + LSTM + Gradient Boosting	92.4	82.4	91.3	90.6	91.0

Table 8
Worst Performance for Specific Cancer Types.

Cancer Type	Model Combination	Main Class Accuracy (%)	Subclass Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Stomach Cancer	VGG19 + RF-XGBoost	78.5	65.0	77.3	76.0	76.6
Brain Cancer	CNN + LSTM + Gradient Boosting	80.2	70.4	79.1	78.0	78.5
Ovarian Cancer	Inception-ResNet + Multi-Scale CNN	82.3	71.5	81.2	80.1	80.6

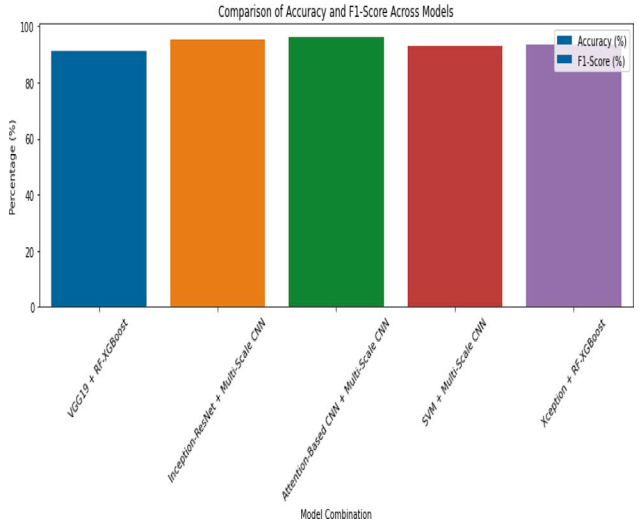

Fig. 5. Accuracy of merged models.

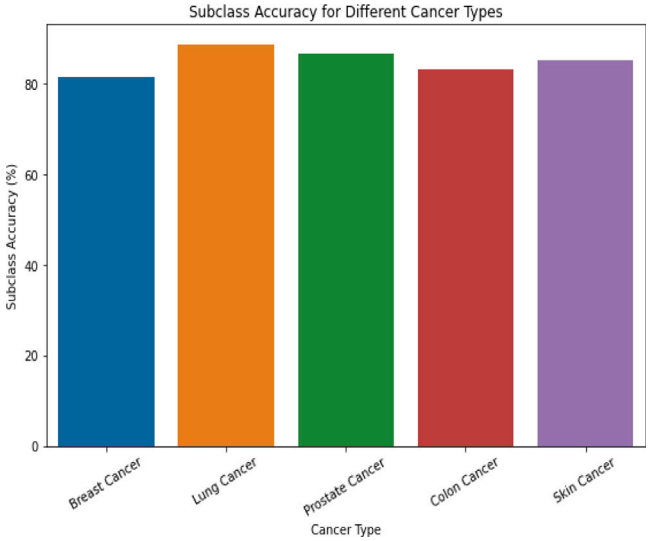

Fig. 6. Accuracy of merged models(subclass).

Specifically, SHAP (SHapley Additive exPlanations) was applied to traditional machine learning models trained on structured data, and Grad-CAM (Gradient-weighted Class Activation Mapping) was used for convolutional neural networks trained on histopathological images.

SHAP for structured data (genomics and clinical)

The SHAP framework was used to analyze feature contributions for models such as XGBoost and Random Forest, which were trained in genomic and clinical metadata. SHAP summary graphs and force graphs provided interpretability for key predictors, such as gene mutations, tumor grade, and patient age, offering transparency in the ML decision-making process.

Grad-CAM for CNN interpretability

Grad-CAM was implemented to visualize the importance of spatial features for CNN models such as VGG19, Inception-ResNet, and Multiscale CNN. These saliency maps highlighted regions in histopathology images that were critical to subtype predictions. The visualizations aligned with clinical features such as nuclear atypia and tissue architecture.

These explainability techniques not only support the validity of model predictions, but also enable clinicians to understand and verify AI-based diagnostic suggestions, making the approach suitable for responsible deployment in real-world clinical settings.

5. Conclusion

We evaluated a variety of deep learning models and combinations of them in classifying cancer type and subtype. The results show that by combining models, AttentionBased CNN, MultiScale CNN, Inception-ResNet, and RFXGBoost resulted in significant performance gains in both main class and subclass identification. In these ensemble methods, we achieve higher precision, precision, recall, and F1 score, demonstrating possible applications to complex and nuanced diagnostic tasks in cancer diagnosis.

Among Attention-Based models, they showed outperforming results, particularly for subclass information which heavily relies on identifying subtle differences between types of cancer. For example, taking into account fine-grained features, Attention-Based CNN + Multi Scale shows a high F1 Score of 94.6%. Similarly, the accuracy of the main class showed that Inception-ResNet + RF – XGBoost consistently performed well with an F1 score of 93. 7%, indicating that feature extraction and regression classification work well.

Although these successes do occur, there are certain model combinations like VGG19 + RF-XGBoost that are seemingly unable to produce good subclass testing accuracy on certain types of cancer (e.g., stomach cancer, brain cancer). But this also stresses out the need to fine-tune model architecture and select the best feature extraction suitable cancer type. Furthermore, models, such as SVM + Multiscale

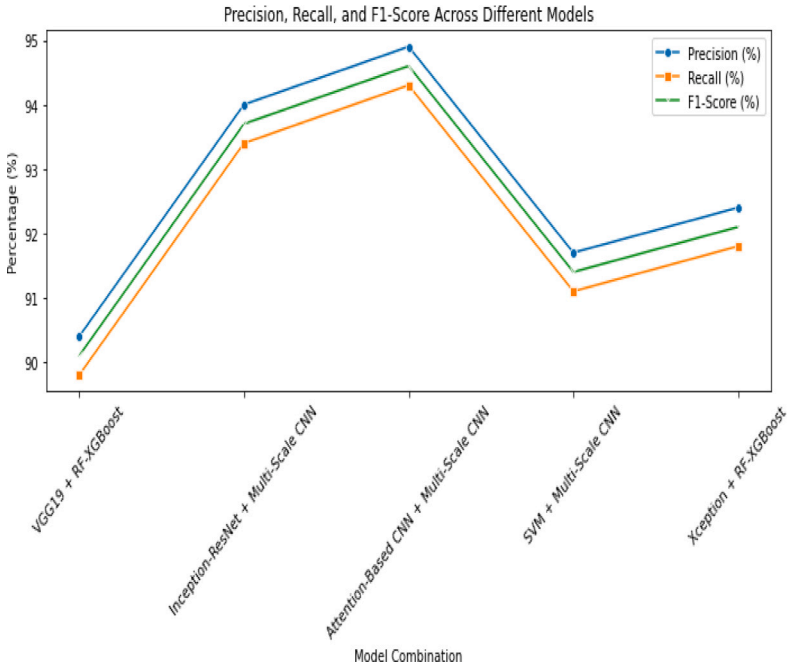

Fig. 7. Precision, Recall, F1-Score.

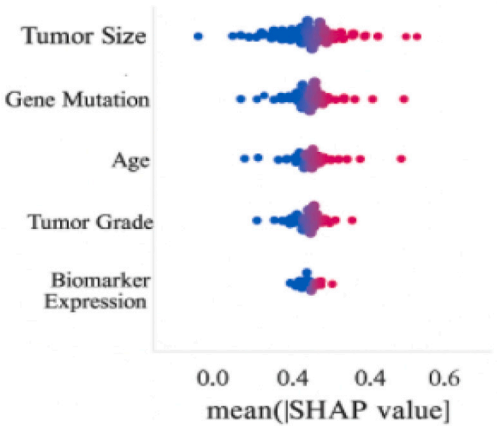

Fig. 8. SHAP summary plot showing the impact of structured features on cancer subtype classification.

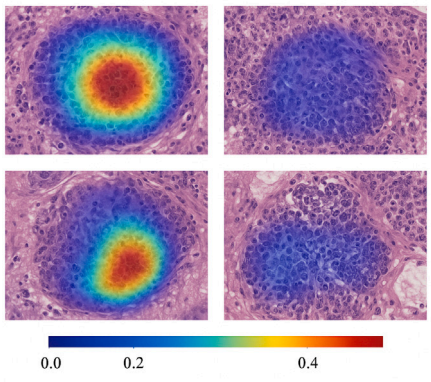

Fig. 9. Sample Grad-CAM overlays highlighting discriminative regions in cancer image classification.

CNN, showed stable but moderate performance, as a combination of traditional machine learning and advanced deep learning appears to hold promise.

In conclusion, the study suggests that with its potential to integrate all models and methods, we can better diagnose and subclassify cancer. Using advances made in the area of deep learning and more traditional classifier approaches, these techniques set the stage for more accurate, more comprehensive, and personalized cancer detection systems. These models must continue to be researched and optimized to the point of real-world clinical applications.

Future directions

There are several promising directions for developing advanced machine and deep learning models for cancer type and subtype classification. One critical avenue is to expand the validation to real-world clinical settings. Although the current study uses the Kaggle multi-cancer dataset, future work will include external validation in large clinically curated repositories such as The Cancer Genome Atlas (TCGA) and BioXpress. These data sets offer diverse demographics of patients, pathology types, and molecular data, enhancing the robustness and applicability of the model.

To ensure real-world clinical readiness, future efforts will also focus on cross-institutional benchmarking and domain adaptation techniques to reduce distributional bias across healthcare systems. Collaborations with hospitals and cancer research institutes will facilitate access to proprietary datasets under Institutional Review Board (IRB)-approved protocols.

In addition, real-time deployment and computational efficiency will be prioritized. Efforts will be directed toward optimizing lightweight, edge-deployable models that can function in resource-constrained settings without compromising performance. This involves compressing models and utilizing architectures that support fast inference, enabling their use in point-of-care diagnostic applications.

Another key direction is improving interpretability and collaboration between clinicians. Human-in-the-loop feedback mechanisms will be integrated into the diagnostic pipeline, allowing oncologists to guide model refinement and validate predictions. This will be supported by continuous development of explainable AI tools, ensuring the transparency of decision-making processes.

Finally, the framework will be extended to support rare cancer types and underrepresented subpopulations, improving fairness and inclusion in model development. By combining these strategies, the goal is to build a clinically viable, generalizable, and ethical AI-based diagnostic system for precision oncology.

CRediT authorship contribution statement

Singamaneni Krishnapriya: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Hyma Birudaraju:** Formal analysis. **M. Madhulatha:** Writing – review & editing. **S. Nagajyothi:** Formal analysis. **K.S. Ranadheer Kumar:** Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Fusing pre-trained convolutional neural networks features for multi-classification of liver cancer data. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC9066403/.
[2] A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe. arXiv. https://arxiv.org/abs/2307.14361.
[3] Transformative breast cancer diagnosis using CNNs with preprocessing. SpringerLink. https://link.springer.com/article/10.1007/s44196-023-00397-1.
[4] A hybrid deep CNN model for brain tumor image multi-classification. BMC Medical Imaging. https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-024-01195-7.
[5] Transfer learning with ensembles of deep neural networks for skin cancer detection in imbalanced data sets. arXiv. https://arxiv.org/abs/2103.12068.
[6] Deep learning in cancer diagnosis and prognosis prediction: A survey. Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1155/2021/9025470.
[7] Exploring the interplay between colorectal cancer subtypes genomic variants and cellular morphology: A deep-learning approach. arXiv. https://arxiv.org/abs/2303.14703.
[8] O.S. Naren, Multi cancer dataset: 8 types of cancer images, 2022, [Online]. Available: https://www.kaggle.com/datasets/obulisainaren/multi-cancer.
[9] Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features. arXiv. https://arxiv.org/abs/1809.02333.